3.1 RESEARCH DESIGN

The research design is the conceptual structure within which research is conducted; it constitutes the blueprint for the collection, measurement and analysis of data. It can be seen at a glance for this study with the help of Chart-3.1.

3.1.1 RESEARCH METHOD-

Explanatory Research Method is used in this study. Causal research, also known as explanatory research is conducted in order to identify the extent and nature of cause-and-effect relationships. Causal research can be conducted in order to assess impacts of specific changes on existing norms, various processes etc. Causal studies focus on an analysis of a situation or a specific problem to explain the patterns of relationships between variables. Experiments are the most popular primary data collection methods in studies with causal research design. Advantages of causal research are follows-

- Causal studies may play an instrumental role in terms of identifying reasons behind a wide range of processes, as well as, assessing the impacts of changes on existing norms, processes etc.
- Causal studies usually offer the advantages of replication if necessity arises.
- These types of studies are associated with greater levels of internal validity due to systematic selection of subjects.

3.1.2 SAMPLING DESIGN

A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sample design may as well lay down the number of items to be included
in the sample i.e., the size of the sample. This study is based on three multi-stage sampling that is a development of the principle of cluster sampling. Chart-3.2 shows the design of sampling of this study.

**CHART-3.1: RESEARCH DESIGN OF THIS STUDY**
UNIT IS SELECTED BY LOTTERY METHOD OF PARAMETRIC SAMPLING
3.1.2.1 UNIVERSE AND SAMPLE OF THE RESEARCH - Registered Women entrepreneurs/enterprises of various district of Allahabad Division namely Allahabad District, Fatehpur District, Pratapgarh District and Kaushambi District are universe of this study. Sample is a part of a population or universe. It means some Registered Women entrepreneurs/enterprises of various district of Allahabad Division namely Allahabad District, Fatehpur District, Pratapgarh District and Kaushambi District are sample of this study because it gives the same results and conclusion by saving money, time and manpower.

3.1.2.2 UNIVERSE SIZE - 800

3.1.2.3 SAMPLE SIZE - 240 (30% of universe)

3.1.2.4 TYPES OF DATA COLLECTING METHOD - The study is conducted on the basis of both Primary and secondary data.

3.1.2.5 SOURCE OF DATA COLLECTION - Secondary data is collected from the Census of India 2011- District (Allahabad, Pratapgarh, Fatehpur and Kaushambi) Census handbook- Directorate of Census Operations Uttar Pradesh. For primary data collection, the universe data of women enterprises has been made on the basis of information gathered from Zila Udhyog Kendra of district of Allahabad Division namely Allahabad District, Fatehpur District, Pratapgarh District and Kaushambi District. A schedule is prepared for the collection of data by personal Interview method for selected respondents. The schedule is also pre tested and later on the principal investigators (researchers) collected the data from selected respondents.

3.1.2.6 SAMPLING METHOD - Mixed sampling both Purposive and proportionate sampling method is used. This thesis is prepared on the basis of a primary survey of Allahabad division (that is chosen purposively) using multistage proportionate random sampling technique.

3.1.2.7 UNIT OF ANALYSIS - Women respondents that are running business in the given district.
3.1.2.8 STATISTICAL TOOL- Statistical tools-namely Central Tendency-Mean, Mode, Dispersion-Standard Deviation, Variances, Standard Error, Bi-Variate Correlation Coefficient (r) -Karl Pearson and Spearman, Part and partial Correlation Coefficient, $R^2$, Bi-Variate and Multiple Regression Analysis, Autocorrelation coefficient, why autocorrelation? Detection of Autocorrelation and removal or adjustment for Autocorrelation (Durbun Watson test, Box Ljung Test) and they were tested by t-test, chi square test and Anova tests. Trend line, Factor Analysis and Logistic Regression are also used. Line diagram, scattered diagram, Pie diagram and Multi Bar Diagram are used for graphical presentation of the data too.

3.2 BRIEF DESCRIPTION OF USED STATISTICAL TOOL IN THIS STUDY

3.2.1 CENTRAL TENDENCY-MEAN OR ARTHMATIC MEAN AND MODE-

The arithmetic mean is defined as being equal to the numerical values of each and every observation divided by the total number of observation. Symbolically it can be represented as-

$$\bar{X} = \frac{\sum X}{N} \text{ or } \frac{\sum f}{f}$$

Mode is defined as the size of the variable at which the frequency is most concentrated or occurs most frequently.

3.2.2 DISPERSION- SATNDARD DEVIATION AND VARIANCES -

Standard Deviation is defined as the positive square root of the arithmetic mean of the squares of deviation taken from Arithmetic mean. Symbolically it can be represented as-

$$\sigma = \sqrt{\frac{\sum (X-\bar{X})^2}{N}} \text{ or } \frac{\sum f}{f}$$

Variance is the square of Standard Deviation that is represented by $\sigma^2$ or V.
3.2.3 STANDARD ERROR -

The standard deviation of the sampling distribution of statistics such as mean, standard deviation, correlation etc., is termed as Standard Error of mean, standard error of standard deviation, standard error of correlation coefficient respectively. The standard error is used in all the sampling methods to estimate the precision of the estimates as well as for estimating the confidence interval of mean and also in tests of hypothesis such as z-test, t-test etc.

3.2.4 BIVARIATE AND MULTIVARIATE CORRELATION COEFFICIENT - KARL PEARSON, SPEARMAN, PART AND PARTIAL CORRELATION COEFFICIENT, $R^2$ –

We may have a data where two characteristics of each individual will be considered for a group of individuals. Such a data is called a Bivariate Data. While an estimate of the combined influence of two or more variables on the observed (dependent) Variable. The coefficient of multiple correlations is a measure of how well a given variable can be predicted using a linear function of a set of other variables.

In a bivariate distribution if the two variables vary in such way that the change in one are followed by changes in the other, then the variables are said to be correlated. The statistical tool with the help of which the relationship between two or more than two variables is discovered and measured is called correlation. When both the variables change in one direction that is when both increase or decrease the relationship between the two variables is called positive or direct. But when the change in the opposite directions that is one increasing and the other is decreasing, the correlation is negative or inverse. For determining the direction of change average values are taken. Correlation only indicates the degree and direction of relationship between two variables.

3.2.4.1 KARL PEARSON CORRELATION COEFFICIENT - A numerical measure of the correlation coefficient between the two variables X and Y denoted by $r$ as given by Karl Pearson is –

$$r=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2(Y-\bar{Y})^2}}$$
3.2.4.2 SPEARMAN CORRELATION COEFFICIENT - Generally characteristics like honest, beauty, efficient, intelligence etc. are better expressed by allotting ranks such as first, second, third, fourth ……….nth etc. The study of correlation of the characteristics expressed by rank is called correlation. The coefficient that is determined from these ranks is known as Spearman’s Rank Correlation Coefficient. Measure of rank correlation a given by Spearman-

\[ r = 1 - \frac{(6\sum d^2)}{n(n^2-n)} \]

3.2.4.3 PART OR SEMIPARTIAL AND PARTIAL CORRELATION COEFFICIENT - Partial correlation is the measure of association between two variables, while controlling or adjusting the effect of one or more additional variables or We measure individual differences in many things, including cognitive ability, personality, interests & motives, attitudes, and so forth. Many times, we want to know about the influence of one IV on a DV, but one or more other IVs pose an alternative explanation. We would like to hold some third variable constant while examining the relations between X and Y. With assignment we can do this by design. With measures of individual differences, we can do this statistically rather than by manipulation. The basic idea in partial and semi partial correlation is to examine the correlations among residuals (errors of prediction). If we regress variable X on variable Z, then subtract X' from X, we have a residual e. This e will be uncorrelated with Z, so any correlation X shares with another variable Y cannot be due to Z. The formula to compute the partial r from correlations is

\[ r_{1,2,3} = \frac{r_{1,2} - r_{1,3}r_{2,3}}{\sqrt{1 - r_{1,3}^2} \sqrt{1 - r_{2,3}^2}} \]

3.2.4.4 COEFFICIENT OF DETERMINATION (R^2) - It measures the proportion or percentage of the total variation in dependent variable explained by the regression model.
3.2.5 BI-VARIATE AND MULTIPLE REGRESSIONS ANALYSIS-

Regression is the determination of a statistical relationship between two or more variables. In simple or bivariate regression, there are only two variables, one variable (defined as independent-X) is the cause of the behavior of another one (defined as dependent variable-Y). Regression can only interpret what exists variable. The basic relationship between X and Y is given by –

\[ \hat{Y} = a + bX \]

Where, the symbol \( \hat{Y} \) denotes the estimated value of Y for a given value of X. This equation is known as the regression equation of Y on X which means that each unit change in X produces a change of b in Y, which is positive for direct and negative for inverse relationships.

When there are two or more than two independent variables, the analysis concerning relationships known as multiple correlations and the equation describing such relationship as the multiple regression equation. Multiple regression equation assumes the form-

\[ \hat{Y} = a + b_1X_1 + b_2X_2 \]

3.2.6 LOGISTIC REGRESSION-

In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. This article covers the case of a binary dependent variable—that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. In the terminology of economics, logistic regression is an example of a qualitative response/discrete choice model. Binary logistic regression is used to predict the odds of being a case based on the values of the independent variables (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a non case. Symbolically we can write it as follows-

\[ P_i = E \left(Y_i = \frac{1}{X_i}\right) = \frac{1}{1 + e^{-\beta}} \]
Where

\[ z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \varepsilon \]

The error term is not observed, and so the is also an unobservable, hence termed "latent".

3.2.7 TREND LINE-

A trend line is an analytical tool used most often in conjunction with a scatter plot (a two-dimensional graph of ordered pairs) to see if there is a relationship between two variables. In other words Trend estimation is a statistical technique to aid interpretation of data. When a series of measurements of a process are treated as a time series, trend estimation can be used to make and justify statements about tendencies in the data, by relating the measurements to the times at which they occurred. This model can then be used to describe the behavior of the observed data. In particular, it may be useful to determine if measurements exhibit an increasing or decreasing trend which is statistically distinguished from random behavior.

3.2.8 t-TEST-

A t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. It can be used to determine if two sets of data are significantly different from each other. Here it is used for testing correlation coefficient of two sets of data. The value of the t statistic is-

\[ t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} \]

3.2.9 CHI SQUARE TEST-

Pearson's chi-squared test ($\chi^2$) is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. It is suitable for unpaired data from large samples. It tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The events considered must be
mutually exclusive and have total probability 1. A common case for this is where the events each cover an outcome of a categorical variable. The value of the test-statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

The chi-squared statistic can then be used to calculate a p-value by comparing the value of the statistic to a chi-squared distribution. The number of degrees of freedom is equal to the number of cells n, minus the reduction in degrees of freedom, p. The result about the numbers of degrees of freedom is valid when the original data are multinomial and hence the estimated parameters are efficient for minimizing the chi-squared statistic.

### 3.2.10 ANOVA TESTS

**Analysis of variance (ANOVA)** is a collection of statistical models used to analyze the differences among group means and their associated procedures (such as "variation" among and between groups), developed by statistician and evolutionary biologist Ronald Fisher. The treatment variance is based on the deviations of treatment means from the grand mean, the result being multiplied by the number of observations in each treatment to account for the difference between the variance of observations and the variance of means. The fundamental technique is a partitioning of the total sum of squares SS into components related to the effects used in the model. For example, the model for a simplified ANOVA with one type of treatment at different levels.

$$SS_{Total} = SS_{Error} + SS_{Treatment}$$

The number of degrees of freedom $DF$ can be partitioned in a similar way: one of these components (that for error) specify a chi-squared distribution which describes the associated sum of squares, while the same is true for "treatments" if there is no treatment effect.

$$DF_{Total} = DF_{Error} + DF_{Treatment}$$
3.2.11 AUTOCORRELATION -

In statistics, the autocorrelation of a random process is the Pearson correlation between values of the process at different times, as a function of the two times or of the time lag. The autocorrelation can be expressed as a function of the time-lag, and that this would be an even function of the lag $\tau = s - t$. This gives the more familiar form

$$R(\tau) = \frac{E[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2},$$

3.2.12 DURBIN WATSON TEST -

In statistics, the Durbin–Watson statistic is a test statistic used to detect the presence of autocorrelation (a relationship between values separated from each other by a given time lag) in the residuals (prediction errors) from a regression analysis. It is named after James Durbin and Geoffrey Watson. If $e_t$ is the residual associated with the observation at time $t$, then the test statistic is

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2},$$

where $T$ is the number of observations. Note that if one has a lengthy sample, then this can be linearly mapped to the Pearson correlation of the time-series data with its lags.\(^2\) Since $d$ is approximately equal to $2(1 - r)$, where $r$ is the sample autocorrelation of the residuals,\(^3\) $d = 2$ indicates no autocorrelation. The value of $d$ always lies between 0 and 4. If the Durbin–Watson statistic is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if Durbin–Watson is less than 1.0, there may be cause for alarm. Small values of $d$ indicate successive error terms are, on average, close in value to one another, or positively correlated. If $d > 2$, successive error terms are, on average, much different in value from one another, i.e., negatively correlated. In regressions, this can imply an underestimation of the level of statistical significance.
To test for **positive autocorrelation** at significance $\alpha$, the test statistic $d$ is compared to lower and upper critical values ($d_{L,\alpha}$ and $d_{U,\alpha}$):

- If $d < d_{L,\alpha}$, there is statistical evidence that the error terms are positively autocorrelated.
- If $d > d_{U,\alpha}$, there is **no** statistical evidence that the error terms are positively autocorrelated.
- If $d_{L,\alpha} < d < d_{U,\alpha}$, the test is inconclusive.
- Positive serial correlation is serial correlation in which a positive error for one observation increases the chances of a positive error for another observation.

To test for **negative autocorrelation** at significance $\alpha$, the test statistic $(4 - d)$ is compared to lower and upper critical values ($d_{L,\alpha}$ and $d_{U,\alpha}$):

- If $(4 - d) < d_{L,\alpha}$, there is statistical evidence that the error terms are negatively autocorrelated.
- If $(4 - d) > d_{U,\alpha}$, there is **no** statistical evidence that the error terms are negatively autocorrelated.
- If $d_{L,\alpha} < (4 - d) < d_{U,\alpha}$, the test is inconclusive.

Negative serial correlation implies that a positive error for one observation increases the chance of a negative error for another observation and a negative error for one observation increases the chances of a positive error for another. The critical values, $d_{L,\alpha}$ and $d_{U,\alpha}$, vary by level of significance ($\alpha$), the number of observations, and the number of predictors in the regression equation. Their derivation is complex—statisticians typically obtain them from the appendices of statistical texts.

### 3.2.13 BOX LJUNG TEST -

The **Ljung–Box test** (named for Greta M. Ljung and George E. P. Box) is a type of statistical test of whether any of a group of autocorrelations of a time series are
different from zero. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags, and is therefore a portmanteau test. This test is sometimes known as the Ljung–Box $Q$ test, and it is closely connected to the Box–Pierce test (which is named after George E. P. Box and David A. Pierce). The Ljung–Box test may be defined as:

$H_0$: The data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process).

$H_a$: The data are not independently distributed; they exhibit serial correlation.

The test statistic is:

$$Q = n(n+2) \sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{n-k}$$

where $n$ is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag $k$, and $h$ is the number of lags being tested. Under $H_0$ the statistic $Q$ follows a $\chi^2(h)$. For significance level $\alpha$, the critical region for rejection of the hypothesis of randomness is

$$Q > \chi^2_{1-\alpha,h}$$

where $\chi^2_{1-\alpha,h}$ is the $\alpha$-quantile of the chi-squared distribution with $h$ degrees of freedom.

3.2.14 REMIDIAL MEASURES FOR AUTOCORRELATION-

The First Difference Method is one method for remedial for autocorrelation. It is used when Durbin Watson statistic is less than coefficient of determination ($d<R^2$).

3.2.15 FACTOR ANALYSIS-

Factor analysis is a useful tool for investigating variable relationships for complex concepts such as socioeconomic status, dietary patterns, or psychological scales. It allows researchers to investigate concepts that are not easily measured directly by collapsing a
large number of variables into a few interpretable underlying factors. The key concept of factor analysis is that multiple observed variables have similar patterns of responses because they are all associated with a latent (i.e. not directly measured) variable. Factor Analysis is a commonly used data/variable reduction technique. This multivariate statistical technique is used for three primary reasons:

- Reduce the number of variables, from large to small
- Establish underlying dimensions between measured variables and constructs and
- Provide construct validity evidence

KMO & Bartlett’s Test of Sphericity is a measure of sampling adequacy that is recommended to check the case to variable ratio for the analysis being conducted. In most academic and business studies, KMO & Bartlett’s test play an important role for accepting the sample adequacy.

3.2.16-KAISER-MEYER-OLKIN (KMO) TEST FOR SAMPLING ADEQUACY -

Kaiser-Meyer-Olkin (KMO) Test is a measure of how suited your data is for Factor Analysis. The test measures sampling adequacy for each variable in the model and for the complete model. The statistic is a measure of the proportion of variance among variables that might be common variance. The lower the proportion, the more suited your data is to Factor Analysis. KMO returns values between 0 and 1. A rule of thumb for interpreting the statistic:

- KMO values between 0.8 and 1 indicate the sampling is adequate.

- KMO values less than 0.6 indicate the sampling is not adequate and that remedial action should be taken. Some authors put this value at 0.5, so use your own judgment for values between 0.5 and 0.6.

- KMO Values close to zero means that there are large partial correlations compared to the sum of correlations. In other words, there are widespread correlations which are a large problem for factor analysis.

For reference, Kaiser put the following values on the results:
• 0.00 to 0.49 unacceptable.
• 0.50 to 0.59 miserable.
• 0.60 to 0.69 mediocre.
• 0.70 to 0.79 middling.
• 0.80 to 0.89 meritorious.
• 0.90 to 1.00 marvelous.

Running the Kaiser-Meyer-Olkin (KMO) Test, The formula for the KMO test is:

$$MO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i} \text{v-bar}}$$

3.2.17 CRONBACH’S ALPHA-

Cronbach’s alpha is a measure of internal consistency, that is, how closely related a set of items are as a group. It is considered to be a measure of scale reliability. We show the formula for the standardized Cronbach’s alpha:

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N - 1) \cdot \bar{c}}$$

Here N is equal to the number of items, c-bar is the average inter-item covariance among the items and v-bar equals the average variance. One can see from this formula that if you increase the number of items, you increase Cronbach’s alpha. Additionally, if the average inter-item correlation is low, alpha will be low. As the average inter-item correlation increases, Cronbach’s alpha increases as well (holding the number of items constant). The theoretical value of alpha varies from zero to 1, since it is the ratio of two variances. However, depending on the estimation procedure used, estimates of alpha can take on any value less than or equal to 1, including negative values, although only positive values make sense. Higher values of alpha are more desirable. Some professionals, as a rule of thumb, require a reliability of 0.70 or higher (obtained on a
substantial sample) before they will use an instrument. Another important aspect that needs mention is the Rotated Component Matrix. While deciding how many factors one would analyze is whether a variable might relate to more than one factor. Rotation maximizes high item loadings and minimizes low item loadings, thereby producing a more interpretable and simplified solution. There are two common rotation techniques - orthogonal rotation and oblique rotation. While orthogonal varimax rotation that produces factor structures that are uncorrelated, oblique rotation produces factors that are correlated. Irrespective of the rotation method used, the primary objectives are to provide easier interpretation of results, and produce a solution that is more parsimonious.

★★★★★★★★